Applied Mathematics Letters 56 (2016) 1–9

Contents lists available at ScienceDirect

Applied Mathematics Letters

www.elsevier.com/locate/aml

Aggregate data and the Prohorov Metric Framework: Efficient gradient computation

H.T. Banks*, Jared Catenacci

Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8212, USA

ARTICLE INFO

Article history:
Received 1 December 2015
Accepted 1 December 2015
Available online 8 December 2015

Keywords:
Inverse problems
Least squares estimation
Estimation of probability distributions
Splines
Prohorov metric

ABSTRACT

We discuss efficient methods for computing gradients in inverse problems for estimation of distributions for individual parameters in models where only aggregate or population level data is available. The ideas are illustrated with two examples arising in applications.

1. Introduction

For years even simple population models based on individual models (see, e.g., the Hare–Lynx models [1, p. 30] and the bacterial growth and diffusion models [1, p. 33], [2], [3, p. 139]) have been based on aggregate population level data for parameter estimation and validation. However, with increased interest in uncertainty quantification and recognition that statistical models for the data collection procedures drive uncertainty statements about the parameters in the underlying mathematical models, the interest in determining correct statistical models as part of parameter estimation or inverse problems has grown. Moreover, it is now recognized that aggregate data is widely (and frequently incorrectly) employed to quantify uncertainty in individual models. This occurs in a ubiquitous range of applied problems including food chemistry efforts [4–6], tracking of labeled substances in proliferating cell populations (e.g., Propagons or prion seeds in amyloid growth in yeast [7–11]), as well as structured population models in marine population studies such as those for mosquito fish [12] and shrimp [13]. In such individual models, one has a mathematical model which describes the behavior of one "individual" which is characterized by a single parameter set which must be estimated using population level or aggregate data.

* Corresponding author.
E-mail address: htbanks@ncsu.edu (H.T. Banks).

http://dx.doi.org/10.1016/j.aml.2015.12.004

In a second class of problems (the aggregate model case), the dynamic mathematical models explicitly depend upon a distribution that must be estimated using aggregate data. This is the case in electromagnetic interrogation problems with a distribution of polarization permittivity and relaxation time parameters for molecules [14–17], in HIV cellular models [18,19], and in wave propagation in viscoelastic materials [17,20,21]. Again in these examples, only aggregate data is available to estimate the embedded probability distributions.

One method for such non-parametric estimation problems of a probability measure is through the Prohorov Metric Framework (PMF) [17,22] developed specifically to treat aggregate data problems (for a summary see [17, Chapter 5]). The PMF provides a theoretical and computational framework in which to estimate an unknown probability measure for which the space $\mathcal{P}(\Omega)$ of probability measures over a compact set Ω is approximated by a finite dimensional space $\mathcal{P}^N(\Omega)$ of dimension N. There are many choices for the approximating space $\mathcal{P}^N(\Omega)$; two popular choices involve using a basis of Dirac measures (zero order splines) or piecewise linear splines to approximate the distributions. In this presentation, our goal is to show how the gradient of a least squares objective function can be found in a efficient manner for inverse problems involving the estimation of a probability measure using the PMF.

2. Problem framework

We assume to have a mathematical model for a dynamical system which is dependent upon a probability measure G as well as Euclidean parameters $\mathbf{q} \in \mathcal{Q}$. We assume that the solution to this system can be obtained either analytically or numerically and denote the solution as $u(x, t; G, \mathbf{q})$. Furthermore we assume that we have a set of observations

$$y_j = u(x_j, t_j; G_0, \mathbf{q}_0) + \epsilon_j, \quad j = 1, \dots, n,$$

where G_0 and $\mathbf{q}_0$ are the true or nominal probability measure and parameters, respectively, and ϵ_j is a realization of the measurement error in the observation process.

Given a set of observations y_j at the points (x_j, t_j), $j = 1, \dots, n$, we would like to estimate the unknown parameters $\mathbf{q} \in \mathcal{Q} \subset \mathbb{R}^\kappa$ and the unknown distribution $G(\theta) \in \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the set of admissible probability measures on $\Omega \subset \mathbb{R}$. Thus, we would like to solve

$$(G, \mathbf{q}) = \underset{(G,\mathbf{q}) \in (\mathcal{P}(\Omega) \times \mathcal{Q})}{\arg\min} J(G, \mathbf{q}), \tag{2.1}$$

where

$$J(G, \mathbf{q}) = \sum_{j=1}^{n} \left(y_j - u(t_j, x_j; G, \mathbf{q})\right)^2. \tag{2.2}$$

We note that (2.1) is an infinite-dimensional optimization problem. Thus, we need to approximate the infinite dimensional space $\mathcal{P}(\Omega)$ with a finite dimensional space $\mathcal{P}^N(\Omega)$ in order to have a computationally tractable finite-dimensional optimization problem

$$(\widehat{G}, \widehat{\mathbf{q}}) = \underset{(G,\mathbf{q}) \in (\mathcal{P}^N(\Omega) \times \mathcal{Q})}{\arg\min} J(G, \mathbf{q}). \tag{2.3}$$

We will consider two finite-dimensional spaces, $\mathcal{P}_D^N(\Omega)$ and $\mathcal{P}_S^N(\Omega)$, to approximate $\mathcal{P}(\Omega)$. The space $\mathcal{P}_D^N$ involves the use of Dirac measures, and the space $\mathcal{P}_S^N$ involves the use of piecewise linear splines. We define these two spaces as

$$\mathcal{P}_D^N(\Omega) = \left\{ G \in \mathcal{P}(\Omega) \;\middle|\; G = \sum_{m=1}^{N} \alpha_m \Delta_{z_m}, \text{ where } \alpha_m \geq 0 \text{ and } \sum_{m=1}^{N} \alpha_m = 1 \right\}, \tag{2.4}$$